The Semantic Integration of Neural Science Data and a Basic Neuroscience Researcher Use Case

March 13, 2006

Author: Donald Doherty mailto:donald.doherty@brainstage.com
Brainstage Research, Inc.

Motivation for the Semantic Integration of Neural Sciences Data

Many major stakeholders in the life sciences and in the neurosciences in particular are motivated to vertically integrate data at the semantic level. Here's a small sample:

- The National Institutes of Health (NIH) see it as essential for an increased return on their nearly $30 billion a year investment in biomedical research. For instance, see the NIH Blueprint for Neuroscience Research (http://neuroscienceblueprint.nih.gov).
- The NIH's neuroinformatics project dubbed The Human Brain Project (http://www.nimh.nih.gov/neuroinformatics/index.cfm) has been specifically aimed at vertical integration of neurosciences data. Stephen H. Koslow led this project and was just snapped up last year by the Allen Institute. (The book "Databasing the Brain" (2005) was edited by Stephen Koslow and Shankar Subramaniam.)
- The Allen Institute for Brain Science (http://www.alleninstitute.org) was recently kicked off by Microsoft co-founder Paul Allen. Their first focus is the Allen Brain Atlas Project (http://www.brainatlas.org).
- The Society for Neurosciences has a Neuroinformatics Committee and the Neuroscience Database Gateway (http://web.sfn.org/content/Programs/NeuroscienceDatabaseGateway/).
- The Biomedical Informatics Research Network (BIRN; http://www.nbirn.net) is focusing on neuroinformatics and semantic integration in the areas of schizophrenia, Alzheimer's disease, and Parkinson's disease among others.

The following aren't necessarily focused on the brain but they are taking on the big picture that inevitably includes the brain:

- Integrative Biology (http://www.integrativebiology.ac.uk/) part of European Union's e-Science project.
- The National Cancer Institute and caBIG.
- The Physiome Project (http://www.physiome.org/)

Motivation to Adopt the W3C Semantic Web Technologies

We all agree that the motivation to semantically integrate neuroscience (and other health and life sciences) data is alive and well amongst stakeholders. The question seems to be how to motivate the stakeholders to use the W3C Semantic Web standards to enable vertical data integration underlying the neurosciences. There are some competing

standards but I believe that all of the stakeholders mentioned above would like to utilize the W3C Semantic Web technologies. What is holding them back?

What the Current W3C Standards Don't Help With

Okay, say the neuroscientist knows to use XML, RDF, and OWL. They're sitting at their bench with their computer loaded with recently gathered data. Now what?

A gene sequencer knows exactly what to publish and in what format. With proteins the question becomes more complicated but at least there are only 21 amino acids and they are aligned in a sequence. (Two and three-dimensional protein structure and other interactions and structural considerations add the complications; see BIND.) Intracellular signaling becomes even more complicated. It is from this level and higher that the data publishing standards are sparse to nonexistent.

The Neuroscientist Use Case

Ask nearly any bench neuroscience researcher what they most need and they will talk about the X number of decades worth of data accumulated in all types of formats, processed in all manner of ways, that they'd like to be able to access, reanalyze, and use moving forward.

They need to be able to access and utilize their own data before they even think about integrating data from others!

A Diverse Community

The neuroscience community is composed of a large number of actors that you might think work closely together but often do not. For example, the basic neuroscience researchers (typically with Ph.D.s) often don't work much with the clinically oriented neuroscience researchers (typically with M.D. s). Their questions, publications, and cultures are often surprisingly different.

There are many other actors in the neuroscience arena with vastly different interests such as the psychologist, clinical practitioner, drug discovery and development people, research student, medical student, nurse, patient, and general public just to name a few.

Semantically enabling genome, proteome, physiome, clinical, and other data will hopefully help to bring better communication to this diverse community.

In the context of our present discussion, use cases are highly dependent on which actors you are talking about. The neuroscientist use case above is from basic neuroscience researchers. It may apply to others such as clinical researchers but I've only asked the basic neuroscientists.

Conclusion

Organizations that put a great deal of money into research and others who realize the benefits of being able to effectively access research results are highly motivated to apply

Semantic Web technologies to neuroscience data integration. In contrast, the neuroscientist generating the results is focused on doing that very difficult task. If they ever even thought of trying to federate their results with others, they dismissed the thought long ago as an unattainable dream.

The problem the basic neuroscientist has is that it's not even clear how to store data in the laboratory in a way that it may be meaningfully accessed and used again by the same people in the same laboratory!

The NIH's Human Brain Project, which was inspired by the large sets of publicly accessible data generated from the Human Genome Project, spent the past decade grappling with this issue which remains the major barrier to the publication of data (beyond genes and proteins) to the Web.

Others are also working on the problem including C. Forbes Dewey, Jr. at MIT. Interestingly, I just did a search for his ExperiBase project and came up with a presentation that it looks like he gave to the W3C (http://schiele.mit.edu:8080/experibase/doc/W3C-Pres-20041028.pdf).

The standards that are used to publish data should be open and in the public domain. Can the HCLSIG help?